

Effect of Reward Function Choices in MDPs with Value-at-Risk

Shuai Ma* Jiayuan Yu*

Abstract

This paper studies Value-at-Risk (VaR) problems in short- and long-horizon Markov decision processes (MDPs) with finite state space and two different reward functions. Firstly we examine the effects of two reward functions under two criteria in a short-horizon MDP. We show that under the VaR criterion, when the original reward function is on both current and next states, the reward simplification will change the VaR. Secondly, for long-horizon MDPs, we estimate the Pareto front of the total reward distribution set with the aid of spectral theory and the central limit theorem. Since the estimation is for a Markov process with the simplified reward function only, we present a transformation algorithm for the Markov process with the original reward function, in order to estimate the Pareto front with an intact total reward distribution.

1 Introduction

A Markov decision process (MDP) is a mathematical framework for formulating the discrete time stochastic control problems. This framework has two features, one is randomness, mainly reflected by transition probability, the other one is controllability, reflected by policy. These two features enable MDP as a natural tool in sequential decision-making for practical problems.

The standard class of optimality criteria concerns the expected total reward, which carries the expectation information of total reward cumulative distribution function (CDF) in several forms, such as the expected discounted total reward in

*Concordia Institute of Information System Engineering, Concordia University, Montréal, Quebec H3G 1M8, Canada (e-mail: m_shua@encs.concordia.ca and jiayuan.yu@concordia.ca).

a finite- or infinite-horizon MDP, average reward in an infinite-horizon MDP, etc. [Derman, 1970, Puterman, 1994].

However, the expectation optimality criteria are not sufficient for many risk-averse problems, where the risk concerns arise not only mathematically but also psychologically. A classic example in psychology is the “St. Petersburg Paradox,” which refers to a lottery with an infinite expected reward, but people only prefer to pay a small amount to play. This problem is thoroughly studied in utility theory, and a recent study brought this idea to reinforcement learning [Prashanth et al., 2015]. A more mathematical example would be autonomous vehicles, in which a sufficient safety factor is more important than a highly expected performance. In general, when high reliability is concerned, the criterion should be formulated as probability instead of expectation.

Two risk criterion classes have been widely examined in recent years. One is the coherent risk measure [Artzner et al., 1998], which has a set of intuitively reasonable properties (convexity, for example). A thorough study in coherent risk optimization can be found in [Ruszczyński and Shapiro, 2006]. The other important class is the mean-variance measure [White, D. J., 1988, Sobel, 1994, Mannor and Tsitsiklis, 2011], in which the expected return is maximized for a given risk level (variance). It is also known as modern portfolio theory.

This paper studies value-at-risk (VaR), which originated from finance. For a given portfolio (an MDP with a policy), a loss threshold (target level), and a time-horizon, VaR concerns the probability that the loss on the portfolio exceeds the threshold over the time horizon. VaR is hard to deal with since it is not a coherent risk measure [Riedel, 2004].

When the criterion concerns the whole distribution instead of the expectation only, the simplification of the reward function will affect the optimal value. For example, the reward function $r' : S \times A \rightarrow \mathbb{R}$ (SA-function) is widely used in many studies on MDP, and it is fine as long as the optimality criterion is presented as an expectation. However, when risk is involved, and the original reward function is $r : S \times A \times S \rightarrow \mathbb{R}$ (SAS-function), the simplification will lead to a non-optimal policy.

In this paper, we study the VaR problems in short and long-horizon MDPs with finite state and action spaces, as well as the effect of the two reward functions. In Section 3, we use a short-horizon MDP to illustrate that under the expected total reward criterion, MDPs with the two reward functions (SA- and SAS-functions) have the same optimal expectation/policy, but different total reward distributions, which result in different VaRs. We also compare the augmented-state 0-1 MDP method and the Pareto front generation for VaR criteria.

Our main contributions are described and discussed in Section 4, which include the following:

- We propose a state-transition transformation algorithm for Markov reward processes derived from MDPs with SAS-functions, in order to estimate the total reward distribution. Since the CDF estimation method is for a Markov process with reward function $r'_\pi : S \rightarrow \mathbb{R}$, the proposed algorithm can transform a Markov process with reward function $r_\pi : S \times S \rightarrow \mathbb{R}$ for the CDF estimation method, and keep the distribution intact.
- We illustrate that both VaR criteria relate to the Pareto front of the total reward distribution set, and we estimate the Pareto front with the aid of spectral theory and the central limit theorem for long-horizon MDPs.

Besides, When the optimality criterion refers to the whole distribution instead of the expectation only, and the original reward function in MDP is $r : S \times A \times S \rightarrow \mathbb{R}$, the reward function should not be simplified. For related studies which concerned VaR or other risk-sensitive criteria, we believe that they should be revisited using our proposed transformation approach instead of the reward simplification.

Related work: This paper adopts the VaR criteria defined in [Filar et al., 1995], which studied the VaR problems on the average reward by separating the state space into communicating and transient classes. Bouakiz and Kebir [Bouakiz and Kebir, 1995] pointed out that the cumulative reward is needed for the VaR criteria, and various properties of the optimality equations were studied in both finite and infinite-horizon MDPs. In a finite-horizon MDP, Wu and Lin [Wu and Lin, 1999] showed that the VaR optimal value functions are target distribution functions, and there exists a deterministic optimal policy. The structure property of optimal policy for an infinite-horizon MDP was also studied. Ohtsubo and Toyonaga [Ohtsubo and Toyonaga, 2002] gave two sufficient conditions for the existence of an optimal policy in infinite-horizon discounted MDPs, and another condition for the unique solution on a transient state set. For the VaR problem with a percentile $\alpha \geq 0.5$, Delage and Mannor [Delage and Mannor, 2010] solved it as a convex “second order cone” program with reward or transition uncertainty. Different from most studies, Boda and Filar [Boda and Filar, 2006] and Kira et al. [Kira et al., 2012] considered the VaR criterion in a multi-epoch setting, in which a risk measure is required to reach an appropriate level at not only the final epoch but also all intermediate epochs.

The VaR problem with a fixed threshold (target value) has been extensively studied. An augmented-state 0-1 MDP was proposed for finite-horizon MDPs

with either integer or real-valued reward functions. The cumulative reward space is included into the state space, and the states which satisfied the threshold was “tagged” by a Boolean reward function. The general reward discretizing error was also bounded [Xu and Mannor, 2011]. In a similar problem named MaxProb MDP, the goal states (in which the threshold is satisfied) were defined as absorbing states, and the problem was solved in a similar way [Kolobov et al., 2011]. Value iteration (VI) was proposed to solve the MaxProb MDP [Yu et al., 1998], and followed by some VI variants. In the topological value iteration (TVI) algorithm, states were separated into strongly-connected groups, and efficiency was improved by solving the state groups sequentially [Dai et al., 2011]. Two methods were presented to separate the states efficiently by integrating depth-first search (TVI-DFS) and dynamic programming (TVI-DP) [Hou et al., 2014]. For both exact and approximated algorithms for VaR with a threshold, the state of the art can be found in [Steinmetz et al., 2016].

Constrained probabilistic MDPs takes VaR as a constraint. The mean-VaR portfolio optimization problem was solved with the Lagrange multiplier for the VaR constraint over a continuous time span [Yiu et al., 2004]. Bonami and Lejeune [Bonami and Lejeune, 2009] solved the mean-variance portfolio optimization problem, and used variants of Chebychev’s inequality to derive convex approximations of the quantile function. Randour et al. [Randour et al., 2015] converted the total discounted reward criterion to an almost-sure percentile problem, and proposed an algorithm based on linear programming to solve the weighted multi-constraint percentile problem. It is also pointed out that randomized policy is necessary when VaR criterion is considered as a constraint, and an example can be found in [Defourny et al., 2008].

2 Preliminaries and Notations

A finite-horizon MDP,

$$\langle N, S, A, r, p, \mu_0, v \rangle,$$

is observed at decision epochs $\{0, 1, \dots, N\}$ and $N < +\infty$; S is a finite state space, and X_t denotes the state at epoch t ; A_x is the legitimate action set associated with each state $x \in S$, $A = \bigcup_{x \in S} A_x$ is a finite action space, and K_t denotes the action at epoch t ; $r : S \times A \times S \rightarrow \mathbb{R}$ is the bounded and measurable reward function, and $r(x, a, y)$ denotes the reward (or cost if negative), given $X_t = x$, $X_{t+1} = y$, $x, y \in S$, and the action $a \in A_x$. This reward function has three arguments, and we name it *SAS-function*; $p(y \mid x, a) = \mathbb{P}(X_{t+1} = y \mid X_t =$

$x, K_t = a)$ denotes the homogeneous transition probability; μ_0 is the initial state distribution; v denotes the salvage function.

The optimal policy π^* is determined by the optimality criterion. A policy π refers to a sequence of decision rules $(\pi_0, \pi_1, \dots, \pi_{N-1})$. Different forms of decision rule are used in different situations, and here we focus on deterministic Markovian decision rules.

In a finite-horizon MDP under an expectation criterion [Puterman, 1994], the SAS-function is usually simplified by

$$r'(x, a) = \sum_{y \in S} r(x, a, y)p(y | x, a), \text{ for all } x \in S, a \in A_x. \quad (1)$$

Here we name the reward function r' *SA-function*. It is suitable to simplify the reward function when the total reward expectation is considered, but when VaR is the criterion, it will lead to a non-optimal result.

2.1 Value-at-Risk Criteria

In this paper, we consider VaR instead of its risk-neutral counterparts. Two VaR problems are considered [Filar et al., 1995]. Denote Π^N as the deterministic policy space with the time horizon N . Given a policy $\pi \in \Pi^N$ and an initial distribution μ_0 , we have the total reward $\Phi_{\mu_0}^\pi = \sum_{t=0}^{N-1} r(X_t, \pi(X_t), X_{t+1}) + v(X_N)$, where $X_0 \sim \mu_0$. To simplify the notation we henceforth denote the total reward by Φ . Denote F_Φ^π as the total reward CDF with a policy π . VaR addresses the following problems.

Problem 1. *Given a percentile $\alpha \in [0, 1]$, find $\rho_\alpha = \sup_{\pi \in \Pi^N} \{\tau \in \mathbb{R} | \mathbb{P}(\Phi > \tau) \geq \alpha\} = \sup_{\pi \in \Pi^N} \{\tau \in \mathbb{R} | F_\Phi^\pi(\tau) \leq 1 - \alpha\}$.*

This problem refers to the quantile function, i.e., $F_\Phi^{\pi^{-1}}$.

Problem 2. *Given a threshold (target level) $\tau \in \mathbb{R}$, find $\eta_\tau = \sup_{\pi \in \Pi^N} \{\alpha \in [0, 1] | F_\Phi^\pi(\tau) \leq 1 - \alpha\}$.*

This problem concerns F_Φ^π . Both VaR problems relate to the Pareto front P_Φ of the CDF set $\{F_\Phi^\pi | \pi \in \Pi^N\}$, i.e., $P_\Phi(x) = \inf_{\pi \in \Pi^N} F_\Phi^\pi(x)$, for all $x \in \mathbb{R}$. As will be illustrated below, when the horizon is short (Section 3), any point along P_Φ is $(\tau, 1 - \eta_\tau)$, and when the horizon is long (Section 4), and every (estimated) F_Φ^π is strictly increasing, any point along P_Φ is $(\rho_\alpha, 1 - \eta_\tau)$. Since

there exists a deterministic optimal policy for finite-horizon MDPs under VaR criteria [Wu and Lin, 1999], we only consider the deterministic policy space.

The SA-function is commonly used in most MDP studies even considering risk ([Filar et al., 1995] for example) instead of the SAS-function. However, under the VaR criteria, if the original reward function is an SAS-function, the simplification will miss the optimality, i.e., neither the policy nor the VaR is optimal. In an MDP with an SAS-function, the simplified SA-function leads to the same optimal policy under the expected total reward criteria, but different optimal policies under the VaR criteria. Here we use a short-horizon inventory control problem to illustrate the effect of reward function on the optimal value under two criteria, and what is VaR about.

3 Short-Horizon MDP for Inventory Problem

The inventory control problem is a straightforward example for illustrating the effect of the two reward functions, since the reward (sales volume) is related to both current and next states. In this short-horizon MDP, we show that under the expected total reward criterion, the simplification (SA-function) of the original reward function (SAS-function) will not affect the optimal value/policy, but change the total reward CDF.

3.1 MDP Description

This example is modified from ([Puterman, 1994], Section 3.2), and the complete problem description can be found in Appendix A. Briefly, the MDP for the short-horizon inventory problem is as follows. The time horizon $N = 2$ ($t \in \{0, 1, 2\}$); the state set $S = \{0, 1, 2\}$ defines all possible inventory levels; the action sets $A_0 = \{0, 1, 2\}$, $A_1 = \{0, 1\}$, $A_2 = \{0\}$ define all legitimate actions (orders) for each state; The two reward functions and the transition probabilities are illustrated in Figure 1. The labels $a(r(x, a, y), p(y \mid x, a))$ along transitions denote the original SAS-function and the transition probability, and the labels $a(r(x, a))$ in the text boxes near states denote the simplified SA-function.¹ The bold parts are an example which illustrates the difference between the two reward functions. Besides, we set the initial state distribution $\mu_0 = [1, 0, 0]$, and the salvage reward

¹For example, the label $2(0, 0.5)$ below the transition from 0 to 1 means that the reward $r(0, 2, 1) = 0$ and the transition probability is 0.5; the label $2(0)$ in the text box near state 0 means when $X_t = 0$ and $K_t = 2$, the simplified reward $r'(0, 2) = 0$.

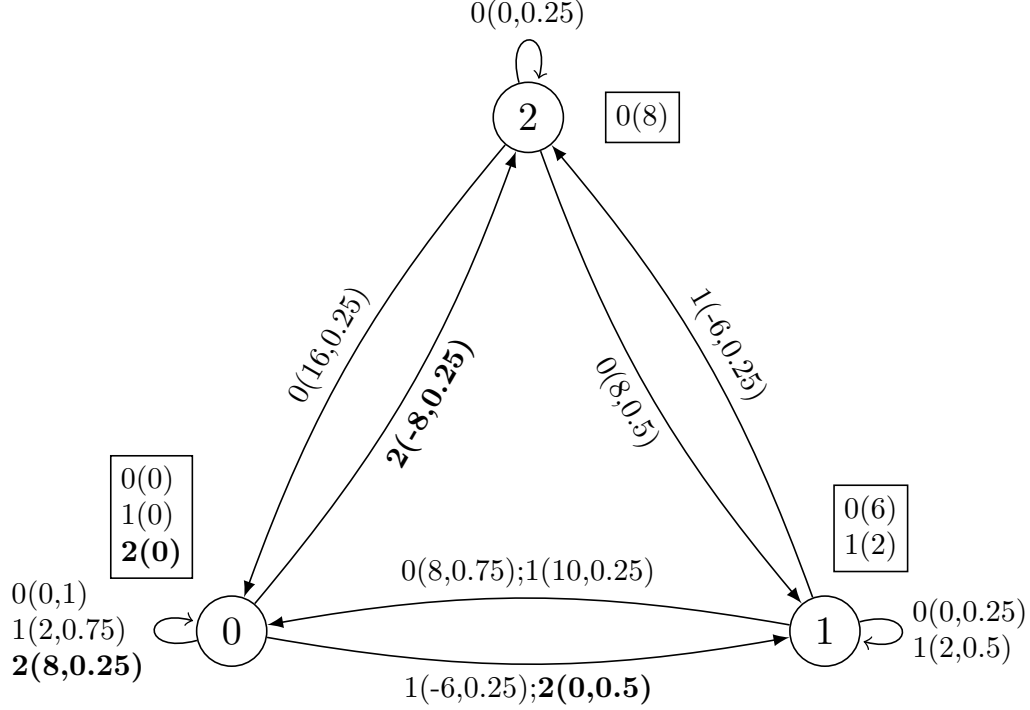


Figure 1: The two reward functions and the transition probabilities for the inventory problem.

$v(x) = x$, for all $x \in S$. Now we have two MDPs with different reward functions: $\langle N, S, A, r, p, \mu_0, v \rangle$ and $\langle N, S, A, r', p, \mu_0, v \rangle$.

3.2 Expected Total Reward Criterion

Under the expected total reward criterion (nominal, discounted or average), the SAS- and SA-functions lead to the same optimal results, but different F_{Φ}^{π} , which results in different VaRs. We illustrate this difference with a short-horizon MDP, and without loss of generality, we consider the nominal expected total reward criterion. The optimal policy for both MDPs is $\pi^* = (\pi_0, \pi_1)$, where $\pi_0(0) = 2$ and $\pi_1(x) = 0$, for $x \in S \setminus \{0\}$. The expected total reward $\mathbb{E}(\Phi) = 6.5625$.

As shown in Figure 2, under the expected total reward criterion, the simplification of SAS-function leads to a different total reward distribution. In the next

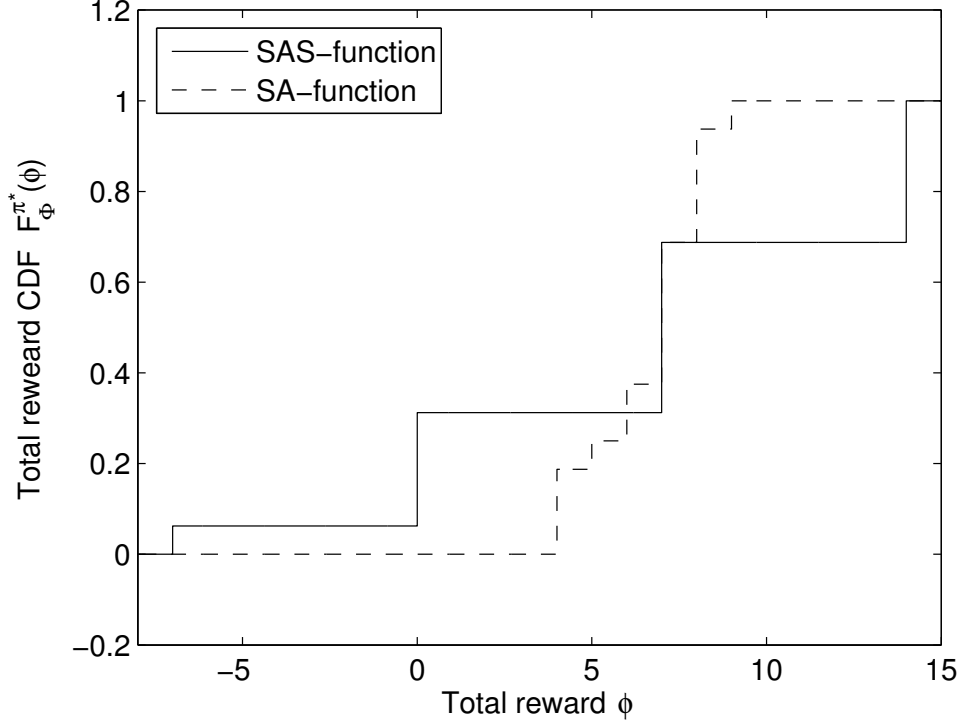


Figure 2: The total reward CDFs of MDPs with SAS- and SA-functions with the same optimal policy π^* for the short-horizon inventory problem under the expected total reward criterion.

section, we discuss the VaR criteria, which refers to the Pareto front of the CDF set, and since the reward simplification changes the total reward distribution, it will miss the optimal VaR.

3.3 VaR Criteria

Unlike the expected total reward criteria, the VaR optimality are not time-consistent, so the backward induction cannot be implemented directly. One method is the augmented-state 0-1 MDP [Xu and Mannor, 2011], which incorporate the cumulative reward space in the state space, and brings in the threshold and reorganizes

the MDP components, in order to calculate the percentile in an expectation way.

3.3.1 Augmented-State 0-1 MDP Method

Since the cumulative reward information is needed for the optimality [Bouakiz and Kebir, 1995], an augmented state space is adopted to keep track of it. For short-horizon MDPs under VaR criterion with a given threshold, Xu and Mannor [Xu and Mannor, 2011] presented a state augmentation method to include the cumulative reward in the state space. Define C as the cumulative reward space, m_a and M_a as the minimum and maximum of the rewards for action a . Then C can be set as $\{0\} \cup_{t=1}^N \bigcup_{a \in A} [t \cdot m_a, t \cdot M_a]$, or we can acquire C by enumerating all possible cumulative reward within a short horizon. Define the augmented state space $S' = S \times C$ for the new MDP. This state augmentation process is also used in several former studies [Bouakiz and Kebir, 1995, Wu and Lin, 1999, Ohtsubo and Toyonaga, 2002, Xu and Mannor, 2011].

For an MDP $\langle N, S', A, r, p, \mu_0, v \rangle^2$, set all reward values to zero and the salvage reward $v' = \mathbb{1}_{[\Phi \geq \tau - v]}$, where τ is the threshold (target level) and Φ is the cumulative reward at the final epoch. This 0-1 MDP enables backward induction to calculate η_τ (defined in VaR Problem 2) as the expectation. Filar et al. [Filar et al., 1995] used the same “0-1” method for infinite-horizon MDPs under both VaR criteria. For all $x, y \in S$ and $x', y' \in S'$, define the action space $A'_{x'} = A_x$ where $x' = (x, \cdot)$; define the transition kernel $p'(y'|x', a) = p(y|x, a)$ where $y' = (y, c_i)$, $x' = (x, c_j)$, $c_i - c_j = r(x, a, y)$ and $c_i, c_j \in C$; define the initial distribution $\mu'_0((x, 0)) = \mu_0(x)$.

Now we have an augmented-state 0-1 MDP $\langle N, S', A, v', p', \mu'_0 \rangle$. Here we proof that the optimal expected total reward of the new MDP equals the solution to VaR Problem 2 in the original MDP, i.e., $\eta_\tau = \sup_{\pi \in \Pi^N} (\mathbb{E}(\Phi))$, and then calculate it with the backward induction.

Lemma 1. *For every finite-horizon MDP, there exists an augmented-state 0-1 MDP, in which the optimal expected total reward equals to the optimal VaR of the original MDP with a threshold $\tau \in \mathbb{R}$.*

Proof. Given an augmented-state 0-1 MDP $\langle N, S', A, v', p', \mu'_0 \rangle$, for all $x' \in S'$, implement the backward induction as follows. **Step 1:** Set $t = N$ and

$$u_N^*(x') = r'_N(x') = \mathbb{1}_{[\Phi \geq \tau - v]}.$$

²Notice that r is original.

Step 2: Set $t = t - 1$, and compute $u_t^*(x')$ by

$$u_t^*(x') = \max_{a \in A_{x'}} \{r'(x', a) + \sum_{y' \in S'} p'(y'|x', a) u_{t+1}^*(y')\},$$

where $r'(x', a) = 0$, therefore,

$$u_t^*(x') = \max_{a \in A_{x'}} \{ \sum_{y' \in S'} p'(y'|x', a) u_{t+1}^*(y') \}.$$

Step 3: If $t = 1$, stop. Otherwise return to Step 2.

Since the only rewards are $r'_N = \mathbb{1}_{[\Phi \geq \tau - v]}$, we have $u_t(x') = P(\Phi \geq \tau | X'_t = x')$, i.e., the probability that the total reward $\Phi \geq \tau$ given any state at any epoch. The optimal policy derived from

$$A_{x',t}^* = \operatorname{argmax}_{a \in A_{x'}} \{P(\Phi \geq \tau | X'_t = x')\}$$

gives the highest probability to reach the threshold. \square

With the help of the new salvage reward function, we are enabled to implement backward induction to compute the corresponding percentile for the VaR criterion with a given threshold. The augmented-state 0-1 MDP (Algorithm 1) is presented as follows.

In the implementation of the algorithm, it is worth noting that, in most instances, it is more efficient to deal with the state space in a time-dependent way, i.e., at each epoch, only a subspace of S' is feasible.

Now we use the augmented-state 0-1 MDP method to solve the inventory control problem described in Section 3.1. We consider the VaR criterion with a given threshold $\tau = 9$. In the MDP with the SAS-function, the optimal policy for the first MDP is $\pi^* = (\pi_0, \pi_1)$, where $\pi_0((0, 0)) = 2$, $\pi_1((0, 2)) = 2$, $\pi_1((0, 8)) = 1$ or 2 , $\pi_1((1, 0)) = 1$ or 2 , $\pi_1((1, 6)) = 0$ or 1 , and $\pi_1((2, -2)) = 0$ or 1 . And the optimal percentile is $\eta_\tau^* = 0.3125$. In the MDP with the SA-function, the optimal policy for the second (simplified) MDP is $\pi^* = (\pi'_0, \pi'_1)$, where $\pi'_0((0, 0)) = 2$, $\pi'_1((2, 0)) = 0$. And the optimal percentile is $\eta_\tau^* = 0.1875$. The conclusion drawn in Section 3.1, which claims that the reward simplification changes the VaR, is verified here.

However, the augmented-state 0-1 MDP method is for VaR Problem 2 with a specified threshold only. In order to achieve all optimal VaRs with any $\tau \in \mathbb{R}$ or

³Section 5.2 in [Xu and Mannor, 2011] described how to convert the policy.

Algorithm 1 Augmented-State 0-1 MDP

Input: a finite-horizon MDP $\langle N, S, A, r, p, \mu_0, v \rangle$ and a threshold $\tau \in \mathbb{R}$.

Output: a deterministic policy π^* and the optimal VaR η_τ .

Generate the cumulative reward set

$$C = \{0\} \bigcup_{t=1}^N \bigcup_{a \in A} [t \cdot m_a, t \cdot M_a],$$

or enumerating all possibilities;

Generate the augmented state space $S' = S \times C$;

Generate the salvage reward $v' = \mathbb{1}_{[\Phi \geq \tau - v]}$;

for all (x', y') **where**

$x' = (x, c_j), y' = (y, c_i), x, y \in S, c_i, c_j \in C$, and $c_i - c_j = r(x, a, y)$ **do**

Construct the transition kernel

$$p'(y' \mid x', a) = p(y \mid x, a);$$

end for

Calculate $\mu'_0(x') = \mu_0(x) \mathbb{1}_{[x'=(x,0)]}$;

Solve the MDP $\langle N, S', A, v', p', \mu'_0 \rangle$ under the expected total reward criterion, and output the policy³.

$\alpha \in [0, 1]$, we can enumerate all the deterministic policies on the augmented state space to acquire the Pareto front P_Φ .

Remark (Pareto front in a short horizon). *Given the Pareto front P_Φ for a short-horizon MDP, for any $\tau \in \mathbb{R}$, $P_\Phi(\tau) = 1 - \eta_\tau$, since*

$$\begin{aligned} 1 - \eta_\tau &= 1 - \sup_{\pi \in \Pi^N} \{\alpha \in [0, 1] \mid F_\Phi^\pi(\tau) \leq 1 - \alpha\} \\ &= \inf_{\pi \in \Pi^N} \{1 - \alpha \in [0, 1] \mid F_\Phi^\pi(\tau) \leq 1 - \alpha\} \\ &= P_\Phi(\tau). \end{aligned}$$

Figure 3 shows the Pareto fronts for MDPs with the two reward functions. It illustrates that the simplification of reward function changes the VaR (Pareto front). Given the two Pareto fronts, we can verify the solution to the VaR problem with a specified threshold $\tau = 9$. Furthermore, for any threshold $\tau \in \mathbb{R}$, we

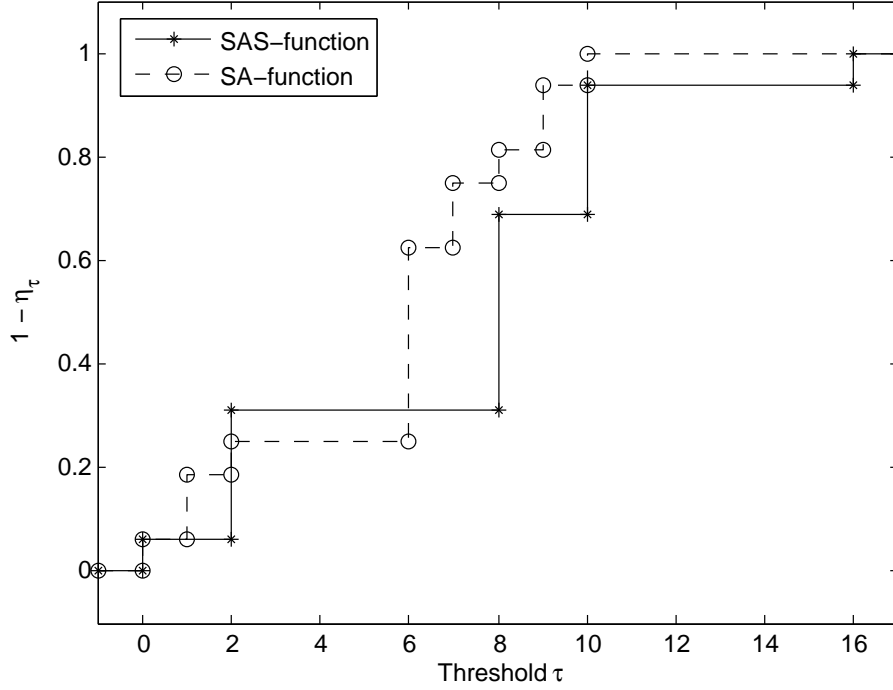


Figure 3: The Pareto fronts for the short-horizon inventory problems with SAS- and SA-functions under some VaR criterion.

can acquire η_τ along the curves.⁴ Table 1 shows the comparison between the two methods.

In conclusion, the reward simplification changes the VaR for a short-horizon MDP. Under the VaR criterion with a threshold, the augmented-state 0-1 MDP method works well for enabling the backward induction algorithm. However, this method fails for long-horizon MDPs, and it works for the VaR problem 2 with a specified threshold only. Since both VaR problems relate to the Pareto front of the total reward CDF set, how to obtain the Pareto front in a short horizon needs further study.

⁴For example, when $\tau = 7.5$, η_τ for the MDP with SAS-function is 0.6875 (1-0.3125), and η'_τ for the MDP with SA-function is 0.25 (1-0.75).

Table 1: Comparison between the augmented-state 0-1 MDP and the Pareto front generation for VaR criteria.

Augmented-state 0-1 MDP	Pareto front generation
Short horizon	Long horizon
Exact result	Estimated result for long-horizon MDPs
VaR Problem 2 only	Both VaR Problems
Backward induction with cumulative reward space	Enumerating stationary policy

4 VaR Criteria in Long-Horizon MDPs

Since it is intractable to find the exact optimal policy for a long-horizon MDP under some the VaR criterion, we look for a deterministic stationary policy instead, i.e., $\pi^* \in \Pi$. With the aid of spectral theory and the central limit theorem, we can estimate the total reward CDF set $\{F_\Phi^\pi\}$ for an MDP with SA-function by enumerating all the deterministic policies. In order to implement the method to MDPs with SAS-functions, we present an algorithm to transform a Markov process with the reward function $r_\pi : S \times S \rightarrow \mathbb{R}$ to one with $r_\pi^\dagger : S^\dagger \rightarrow \mathbb{R}$, where $S^\dagger = S \times S$. This method is for both VaR problems.

4.1 Total Reward CDF Estimation

Firstly we estimate the CDF in a long-horizon Markov reward process derived from an MDP with SA-function. Given an MDP $\langle N, S, A, r', p, \mu_0 \rangle$ ⁵ and a deterministic policy π , we have a Markov reward process $\langle N, S, r'_\pi, p_\pi, \mu_0 \rangle$. For $x, y \in S$, the reward is $r'_\pi(x) = r'(x, \pi(x))$, and the transition kernel is $p_\pi(x, y) = p(x, \pi(x), y)$.

Kontoyiannis and Meyn [Kontoyiannis and Meyn, 2003] proposed a method to estimate F_Φ^π . In a positive recurrent Markov process with invariant probability measure (stationary distribution) ξ , the total reward $\Phi_N = \sum_{t=0}^{N-1} r'_\pi(X_t)$, and the averaged reward $\zeta(r'_\pi) = \lim_{N \rightarrow \infty} \frac{1}{N} \mathbb{E}(\sum_{t=0}^{N-1} r'_\pi(X_t))$, which can be expressed

⁵Though v is ignored when the horizon is long, it can be involved if necessary.

as $\zeta = \xi r'_\pi$. Define the limit $\hat{r}'_\pi = \lim_{N \rightarrow \infty} \mathbb{E}_{\mu_0}(\Phi_N - N\zeta)$, which solves the Poisson equation

$$P\hat{r}'_\pi = \hat{r}'_\pi - r'_\pi + \zeta,$$

where P is the transition matrix and $P(x, y) = p_\pi(x, y)$. Two assumptions ([Kontoyiannis and Meyn, 2003], Section 4) are needed for the CDF estimation.

Assumption 1. *The Markov process X is geometrically ergodic with a Lyapunov function $V : X \rightarrow [1, \infty)$ such that $\zeta(V^2) < \infty$.*

Assumption 2. *The (measurable) function $r'_\pi : S \rightarrow [-1, 1]$ has zero mean and nontrivial asymptotic variance $\sigma^2 = \lim_{N \rightarrow \infty} \text{var}_x[(\Phi_N)/\sqrt{N}]$.*

Under the two assumptions, we show the Edgeworth expansion theorem for nonlattice functionals (Theorem 5.1 in [Kontoyiannis and Meyn, 2003]) as follows.

Theorem 1. *Suppose that X and the strongly nonlattice functional r'_π satisfy Assumptions 1 and 2, and let $G_N(y)$ denote the distribution function of the normalized partial sums $(\Phi_N - N\zeta(r'_\pi))/\sigma\sqrt{N}$:*

$$G_N(y) = \mathbb{P}\{(\Phi_N - N\zeta(r'_\pi))/\sigma\sqrt{N} \leq y\}, \text{ for all } y \in \mathbb{R}.$$

Then, for all $x_0 \in S$ and as $N \rightarrow \infty$,

$$G_N(y) = g(y) + \frac{\gamma(y)}{\sigma\sqrt{N}} \left[\frac{\kappa}{6\sigma^2} (1 - y^2) - \hat{r}'_\pi(x_0) \right] + o(N^{-0.5}),$$

where $\gamma(y)$ denotes the standard normal density, $g(y)$ is the corresponding distribution function, and κ is a constant related to the third moment of Φ_N/\sqrt{N} . The formulae for κ , \hat{r}'_π and σ^2 can be found in Appendix B.

4.2 State-Transition Transformation

For a Markov process derived from an MDP with an SAS-function and a stationary policy π , we cannot apply the method directly since the reward function of the Markov process is $r_\pi : S \times S \rightarrow \mathbb{R}$. If the reward function is simplified by Equation (1), the VaR will be affected as illustrated in Section 3. In order to implement the estimation, we propose a method to transform the Markov process to a Markov process with a reward function $r_\pi^\dagger : S^\dagger \rightarrow \mathbb{R}$ which shares the same F_Φ^π of the original Markov process.

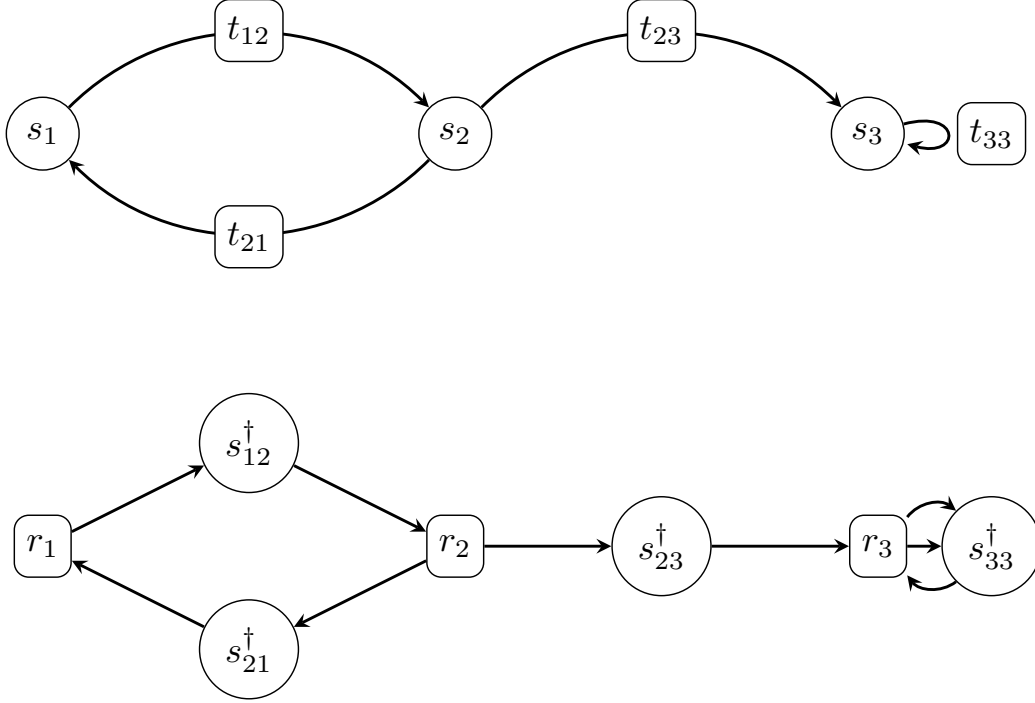


Figure 4: Roles of states and transitions in the transformed Markov process.

Figure 4 illustrates what roles the states and transitions play in the original Markov process (above) and its transformed counterpart (below). In the original Markov process, s_i denotes a state and t_{ij} denotes a transition from s_i to s_j . In a transformed Markov process, the original state s_i becomes a “router” r_i , which connects input nodes (transformed states) s_{ji}^\dagger to output nodes s_{ik}^\dagger .

Since we take state transitions as states, we name this algorithm *state-transition transformation*. It is clear that F_Φ^π for the new Markov process is equivalent to that of the original one. In short, for MDPs with SAS-functions, each stationary policy leads to a Markov process with the reward function r_π , and in order to implement the CDF estimation without simplifying the reward function by Equation (1), we implement the state-transition transformation (Algorithm 2) to generate a Markov process with the same F_Φ^π and a reward function $r_\pi^\dagger : S^\dagger \rightarrow \mathbb{R}$.

In the same MDP setup outlined in Section 3.1, we estimate P_Φ in a long horizon. We set $N = 500$ and implement the state-transition transformation to an MDP with an SAS-function under a stationary policy. In Figure 5, we can see

Algorithm 2 State-Transition Transformation

Input: the original Markov process $\langle N, S, r_\pi, p_\pi, \mu_0 \rangle$, which is derived from the MDP with an SAS-function and a stationary policy π .

Output: a Markov process $\langle N^\dagger, S^\dagger, r_\pi^\dagger, p_\pi^\dagger, \mu_0^\dagger \rangle$ (A salvage reward v^\dagger can be added if necessary).

Set the horizon $N^\dagger = N - 1$;

Generate the state space $S^\dagger = S \times S$;

for all $x^\dagger = (x, y)$ **where** $x, y \in S$ **do**

 Construct the reward function $r_\pi^\dagger(x^\dagger) = r_\pi(x, y)$;

 Construct the transition kernel

$p_\pi^\dagger(x^\dagger | y^\dagger) = p_\pi(y | x)$ where $y^\dagger = (\cdot, x) \in S^\dagger$;

 Set the initial state distribution

$\mu_0^\dagger(x^\dagger) = \mu_0(x)p(y | x)$;

end for

that the simplification of the SAS-function changes the VaR when the horizon is long, and the Pareto front with SAS-function has a wider support than that with SA-function.

Remark (Pareto front in a long horizon). *In a long-horizon MDP, given an estimated P_Φ which is strictly increasing, for any $\alpha \in [0, 1]$, there exists a unique $\tau_0 \in \mathbb{R}$ s.t. $\alpha = 1 - P_\Phi(\tau_0) = \eta_{\tau_0}$, and*

$$\begin{aligned} \rho_\alpha &= \sup_{\pi \in \Pi^N} \{ \tau \in \mathbb{R} | F_\Phi^\pi(\tau) \leq 1 - \alpha \} \\ &= \sup_{\pi \in \Pi^N} \{ \tau \in \mathbb{R} | F_\Phi^\pi(\tau) \leq 1 - \eta_{\tau_0} \} \\ &= \sup_{\pi \in \Pi^N} \{ \tau \in \mathbb{R} | F_\Phi^\pi(\tau) \leq \\ &\quad \inf_{\pi \in \Pi^N} \{ 1 - \alpha \in [0, 1] | F_\Phi^\pi(\tau_0) \leq 1 - \alpha \} \} \\ &= \sup_{\pi \in \Pi^N} \{ \tau \in \mathbb{R} | F_\Phi^\pi(\tau) \leq F_\Phi^\pi(\tau_0) \}. \\ &= \tau_0. \end{aligned}$$

5 Conclusion and Discussions

In this paper, we studied short- and long-horizon MDPs with finite state space under VaR criteria, and the effect of the simplification of reward function. In

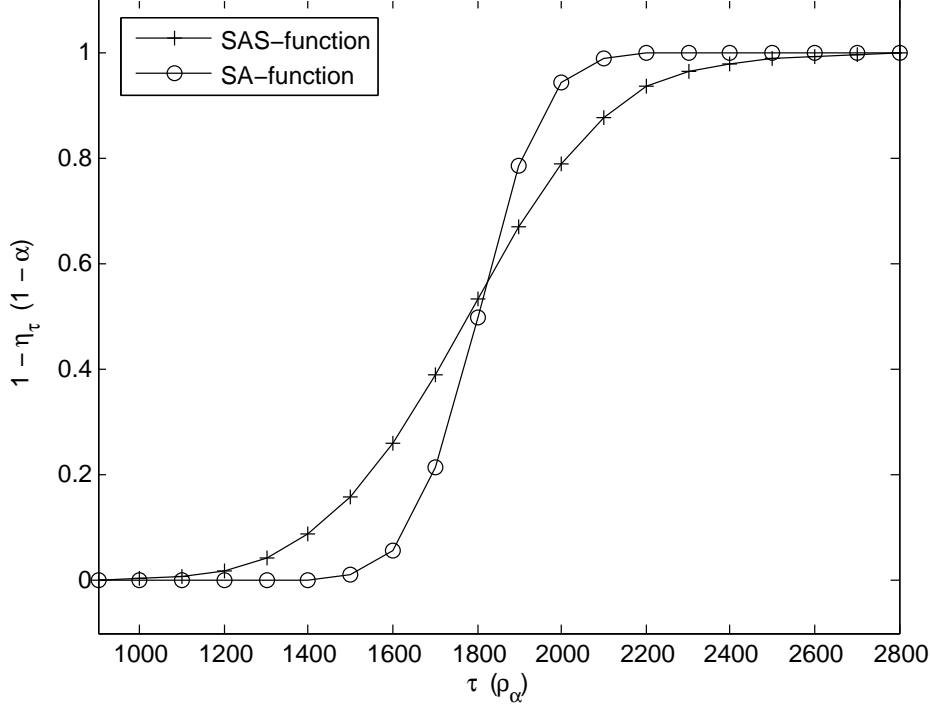


Figure 5: Estimated Pareto fronts for the long-horizon inventory problems with SAS- and SA-functions.

short-horizon MDPs, firstly we illustrated that when the original reward function is an SAS-function, the reward simplification does not affect the optimal value/policy under the expected total reward criterion, but changes the total reward CDF. Secondly we considered VaR criteria, we solved the VaR Problem 2 with the augmented-state 0-1 MDP method in an expectation way, and we enumerated all policies to obtain the Pareto front P_Φ of the total reward CDF set. when the horizon is long, we estimate F_Φ^π for every deterministic policy in order to obtain P_Φ . Since the estimation method is only for Markov processes derived from MDPs with SA-function, we propose a transformation algorithm to make it feasible for the MDPs with SAS-function.

The state-transition transformation enables the original transitions to have

properties as states. Is there a similar transition for MDP, which can convert the MDP with SAS-function to an MDP with SA-function with an equivalent total reward distribution? Two components of MDP will deteriorate if the transformation is applied directly to MDPs. The initial state distribution will be determined by the decision rule at the first epoch, and the salvage reward will be determined by the decision rule at the final epoch. When we concern the long-run performance, and both effects can be ignored, we can implement a similar MDP transformation directly.

VaR concerns the threshold-percentile pair, and the optimality of one comes into conflict with the other as they are virtually non-increasing functions of each other [Filar et al., 1995]. One future study is to estimate the Pareto front without enumerating all the policies. A special case is that there exists an optimal policy π^* , i.e., $F_{\Phi}^{\pi^*}(\tau) = \inf_{\pi \in \Pi^N} F_{\Phi}^{\pi}(\tau)$, for all $\tau \in \mathbb{R}$. Ohtsubo and Toyonaga [Ohtsubo and Toyonaga, 2002] gave two sufficient conditions for the existence of this optimal policy in infinite-horizon discounted MDPs. Another idea is to consider it as a dual-objective optimization. Zheng [Zheng, 2009] studied the dual-objective MDP concerning variance and CVaR, which might provide some insight.

Under a VaR criterion, the simplification of reward function affects the VaR. We believe that some practical problems with respect to VaR should be revisited using our proposed transformation approach when the reward function is an SAS-function.

Appendices

A Inventory Problem Description

Section 3.2.1 in [Puterman, 1994] described the model formulation and some assumptions for a single-product stochastic inventory control problem. Briefly, at t , define X_t as the inventory level before the order, K_t as the order quantity, D_t as the demand with a time-homogeneous probability distribution $\mathbb{P}(D_t = i)$, where $i \in \mathbb{N}$, then we have $X_{t+1} = \max\{X_t + K_t - D_t, 0\}$.

For $u \in \mathbb{N}$, define $c(u)$ as the cost to order u units, and a fixed cost $K \geq 0$ for placing orders, then we have the order cost $O(u) = (K + c(u))\mathbb{1}_{[u>0]}$. $f(u)$ denotes the revenue when u units of demand is fulfilled. Then we have the SAS-function $r(X_t, K_t, X_{t+1}) = f(X_t + K_t - X_{t+1}) - O(K_t)$. Here we ignore the maintenance fee to simplify the problem.

We set the parameters as follows. The time horizon $N = 2$, the fixed order cost $K = 4$, the variable order cost $c(u) = 2u$, the salvage reward $v = X_N$, the warehouse capacity $M = 3$, and the price $f(u) = 8u$. The probabilities of demands are $\mathbb{P}(D_t = 0) = 0.25$, $\mathbb{P}(D_t = 1) = 0.5$, $\mathbb{P}(D_t = 2) = 0.25$ respectively. Initial distribution $\mu_0(0) = 1$, i.e., $X_0 = 0$. Firstly we calculate the SAS-function by $r(X_t, K_t, X_{t+1}) = f(X_t + K_t - X_{t+1}) - O(K_t)$. Secondly, we calculate the SA-function r' by Equation (1). Now we have two MDPs with different reward functions: $\langle N, S, A, r, p, \mu_0, v \rangle$ and $\langle N, S, A, r', p, \mu_0, v \rangle$.

B CDF Estimation for Long-Horizon Markov Reward Process

κ is a constant related to the third moment of the sum $\sum_{t=0}^{N-1} r'(X_t)/\sqrt{N}$. As described in [Kontoyiannis and Meyn, 2003], Section 5 and [Yu et al., 2015], $\kappa = \kappa_1 + \kappa_2 + \kappa_3$, where

$$\kappa_1 = \sum_{x_0 \in S} r'^3(x_0) \mu_0(x_0),$$

$$\kappa_2 = 3 \sum_{i \neq 0}^N \left(\sum_{x_0 \in S} r'^2(x_0) \mu_0(x_0) \times \sum_{x_i \in S} r'(x_i) P^i(x_0, x_i) \right),$$

and

$$\begin{aligned} \kappa_3 = & 6 \sum_{i,j=1}^N \left(\sum_{x_0 \in S} r'(x_0) \mu_0(x_0) \right. \\ & \times \sum_{x_i \in S} r'(x_i) P^i(x_0, x_i) \\ & \times \left. \sum_{x_j, x_{i+j} \in S} r'(x_{i+j}) P^j(x_i, x_j) \right). \end{aligned}$$

As studied in [Glynn and Meyn, 1996] and [Yu et al., 2015], firstly define a kernel $\Xi(x, \cdot) = \xi(\cdot)$, and $H = I - P - \Xi$. Then obtain the fundamental kernel $Z = (H)^{-1}$ if H^{-1} exists. Glynn and Meyn [Glynn and Meyn, 1996] showed that

$$\hat{r}' = Z(r' - \zeta(r')\mathbb{1}).$$

Theorem 17.4.4 in [Meyn and Tweedie, 2009] showed that when $\sigma^2 < \infty$ and $\zeta(r') < \infty$, the asymptotic variance σ^2 can be calculated by

$$\sigma^2 = \sum_{x \in S} [\hat{r}'^2(x) - (P\hat{r}')(x)^2] \xi(x).$$

References

- [Artzner et al., 1998] Artzner, P., Delbaen, F., Eber, J., and Heath, D. (1998). Coherent measures of risk. *Mathematical Finance*, 9(3):1–24.
- [Boda and Filar, 2006] Boda, K. and Filar, J. A. (2006). Time Consistent Dynamic Risk Measures. *Mathematical Methods of Operations Research*, 63(1):169–186.
- [Bonami and Lejeune, 2009] Bonami, P. and Lejeune, M. A. (2009). An Exact Solution Approach for Portfolio Optimization Problems Under Stochastic and Integer Constraints. *Operations Research*, 57(3):650–670.
- [Bouakiz and Kebir, 1995] Bouakiz, M. and Kebir, Y. (1995). Target-level criterion in Markov decision processes. *Journal of Optimization Theory and Applications*, 86(1):1–15.
- [Dai et al., 2011] Dai, P., Weld, D. S., and Goldsmith, J. (2011). Topological value iteration algorithms. *Journal of Artificial Intelligence Research*, 42:181–209.
- [Defourny et al., 2008] Defourny, B., Ernst, D., and Wehenkel, L. (2008). Risk-Aware Decision Making and Dynamic Programming. In *Proceedings of NIPS-08 Workshop on Model Uncertainty and Risk in Reinforcement Learning*, pages 1–8.
- [Delage and Mannor, 2010] Delage, E. and Mannor, S. (2010). Percentile optimization for markov decision processes with parameter uncertainty. *Operations research*, 58(1):203–213.
- [Derman, 1970] Derman, C. (1970). *Finite State Markovian Decision Processes*. Academic Press, Inc.

- [Filar et al., 1995] Filar, J. A., Krass, D., Ross, K. W., and Member, S. (1995). Percentile Performance Criteria For Limiting Average Markov Decision Processes. *IEEE Transactions on Automatic Control*, 40(1):2–10.
- [Glynn and Meyn, 1996] Glynn, P. W. and Meyn, S. P. (1996). A lyapunov bound for solutions of poisson’s equation. *The Annals of Probability*, pages 916–931.
- [Hou et al., 2014] Hou, P., Yeoh, W., and Varakantham, P. (2014). Revisiting risk-sensitive mdps: New algorithms and results. In *Proceedings of the International Conference on Automated Planning and Scheduling (ICAPS)*, pages 136–144.
- [Kira et al., 2012] Kira, A., Ueno, T., and Fujita, T. (2012). Threshold probability of non-terminal type in finite horizon Markov decision processes. *Journal of Mathematical Analysis and Applications*, 386(1):461–472.
- [Kolobov et al., 2011] Kolobov, A., Mausam, Weld, D. S., and Geffner, H. (2011). Heuristic search for generalized stochastic shortest path mdps. In *Proceedings of the International Conference on Automated Planning and Scheduling (ICAPS)*, pages 130–137.
- [Kontoyiannis and Meyn, 2003] Kontoyiannis, I. and Meyn, S. P. (2003). Spectral theory and limit theorems for geometrically ergodic markov processes. *Annals of Applied Probability*, 13:304–362.
- [Mannor and Tsitsiklis, 2011] Mannor, S. and Tsitsiklis, J. (2011). Mean-Variance Optimization in Markov Decision Processes. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pages 1–22.
- [Meyn and Tweedie, 2009] Meyn, S. P. and Tweedie, R. L. (2009). *Markov chains and stochastic stability*. Springer Science & Business Media.
- [Ohtsubo and Toyonaga, 2002] Ohtsubo, Y. and Toyonaga, K. (2002). Optimal policy for minimizing risk models in Markov decision processes. *Journal of mathematical analysis and applications*, 271(1):66–81.
- [Prashanth et al., 2015] Prashanth, L. A., Cheng, J., Fu, M., Marcus, S., and Jun, L. G. (2015). Cumulative Prospect Theory Meets Reinforcement Learning : Estimation and Control. *Working Paper*, pages 1–27.

- [Puterman, 1994] Puterman, M. L. (1994). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley.
- [Randour et al., 2015] Randour, M., Raskin, J., and Sankur, O. (2015). Percentile Queries in Multi-dimensional Markov Decision Processes. *Computer Aided Verification*, 9206:123–139.
- [Riedel, 2004] Riedel, F. (2004). Dynamic coherent risk measures. *Stochastic Processes and their Applications*, 112(2):185–200.
- [Ruszczynski and Shapiro, 2006] Ruszczynski, A. and Shapiro, A. (2006). Optimization of Convex Risk Functions. *Mathematics of Operations Research*, 31(3):433–452.
- [Sobel, 1994] Sobel, M. J. (1994). Mean-Variance Tradeoffs in an Undiscounted MDP. *Operations Research*, 42(1):175–183.
- [Steinmetz et al., 2016] Steinmetz, M., Hoffmann, J., and Buffet, O. (2016). Goal probability analysis in mdp probabilistic planning: Exploring and enhancing the state of the art. *Journal of Artificial Intelligence Research*, 57:229–271.
- [White, D. J., 1988] White, D. J. (1988). Mean , Variance , and Probabilistic Criteria in Finite Markov Decision Processes : A Review. *Journal of Optimization Theory and Applications*, 56(1):1–29.
- [Wu and Lin, 1999] Wu, C. and Lin, Y. (1999). Minimizing Risk models in Markov decision process with policies depending on target values. *Journal of Mathematical Analysis and Applications*, 23(1):47–67.
- [Xu and Mannor, 2011] Xu, H. and Mannor, S. (2011). Probabilistic goal Markov decision processes. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2046–2052.
- [Yiu et al., 2004] Yiu, K. F. C., Wang, S. Y., and Mak, K. L. (2004). Optimal portfolios under a value-at-risk constraint. *Journal of Economic Dynamics and Control*, 28(7):1317–1334.
- [Yu et al., 2015] Yu, P., Yu, J. Y., and Xu, H. (2015). Central-limit approach to risk-aware markov decision processes. *arXiv:1512.00583*.

- [Yu et al., 1998] Yu, S. X., Lin, Y., and Yan, P. (1998). Optimization models for the first arrival target distribution function in discrete time. *Journal of mathematical analysis and applications*, 225(1):193–223.
- [Zheng, 2009] Zheng, H. (2009). Efficient frontier of utility and CVaR. *Mathematical Methods of Operations Research*, 70(1):129–148.